remote sensing

MDPI

Article

Ready-to-Use Methods for the Detection of Clouds, Cirrus, Snow, Shadow, Water and Clear Sky Pixels in Sentinel-2 MSI Images

André Hollstein *, Karl Segl, Luis Guanter, Maximilian Brell and Marta Enesco

Helmholtz-Zentrum Potsdam, Deutsches GeoForschungsZentrum GFZ, Telegrafenberg, 14473 Potsdam, Germany; karl.segl@gfz-potsdam.de (K.S.); luis.guanter@gfz-potsdam.de (L.G.); maximilian.brell@gfz-potsdam.de (M.B.); marta.enesco3@gfz-potsdam.de (M.E.)
* Correspondence: andre.hollstein@gfz-potsdam.de; Tel.: +49-331-288-28969

Academic Editors: Clement Atzberger and Prasad S. Thenkabail
Received: 27 April 2016; Accepted: 1 August 2016; Published: 18 August 2016

Abstract: Classification of clouds, cirrus, snow, shadows and clear sky areas is a crucial step in the pre-processing of optical remote sensing images and is a valuable input for their atmospheric correction. The Multi-Spectral Imager on board the Sentinel-2's of the Copernicus program offers optimized bands for this task and delivers unprecedented amounts of data regarding spatial sampling, global coverage, spectral coverage, and repetition rate. Efficient algorithms are needed to process, or possibly reprocess, those big amounts of data. Techniques based on top-of-atmosphere reflectance spectra for single-pixels without exploitation of external data or spatial context offer the largest potential for parallel data processing and highly optimized processing throughput. Such algorithms can be seen as a baseline for possible trade-offs in processing performance when the application of more sophisticated methods is discussed. We present several ready-to-use classification algorithms which are all based on a publicly available database of manually classified Sentinel-2A images. These algorithms are based on commonly used and newly developed machine learning techniques which drastically reduce the amount of time needed to update the algorithms when new images are added to the database. Several ready-to-use decision trees are presented which allow to correctly label about 91% of the spectra within a validation dataset. While decision trees are simple to implement and easy to understand, they offer only limited classification skill. It improves to 98% when the presented algorithm based on the classical Bayesian method is applied. This method has only recently been used for this task and shows excellent performance concerning classification skill and processing performance. A comparison of the presented algorithms with other commonly used techniques such as random forests, stochastic gradient descent, or support vector machines is also given. Especially random forests and support vector machines show similar classification skill as the classical Bayesian method.

Keywords: Sentinel-2 MSI; cloud detection; snow detection; cirrus detection; shadow detection; Bayesian classification; machine learning; decision trees

1. Introduction

The detection of clouds, cirrus, and shadows is among the first processing steps after processing raw instrument measurements to at-sensor radiance or reflectance values. A robust discrimination of cloudy, cirrus-contaminated, and clear sky pixels is crucial for many applications, including the retrieval of surface reflectance within atmospheric correction (e.g., see [1,2]) or the co-registration with other images (e.g., see [3,4]). The retrieval of surface reflection becomes impossible for optically thick clouds and pixels affected by cirrus and shadows must be treated as individual cases for a physically correct retrieval. Many applications benefit if a detection of snow and water is additionally performed (e.g., see [5,6]). In that respect, such a classification is an essential pre-processing step before higher-level algorithms can be applied (e.g., see [7,8]). Examples are the application of agriculture-related products which might require clear sky pixels as input (e.g., see [9,10]).

Here we describe ready-to-use classification methods which are applicable for the series of Sentinel-2 MSI (Multi-Spectral Imager) [11–13] instruments. Ready-to-use methods are in a state where their application by a user requires no further research and only a little work for initial setup. The first Sentinel-2 in a series of at least four was launched in 2015, became operational in early 2016, and the Copernicus program aims at having two operational instruments in orbit at a time until 2020. The MSI offers optimized bands for Earth-observation applications as well as for the detection of visible and sub-visible cirrus clouds for which the so-called cirrus channel B10 at 1.38 μm is essential. Such a band is also present it the Operational Land Imager (OLI) [14–16] instrument, which is the most recent installment of the NASA Landsat series. OLI and MSI share similar bands such as the SWIR bands at 1.61 μm and 2.19 μm, but differ in that OLI includes thermal bands while the MSI includes a higher spectral sampling within the red edge. A substantial difference between the two missions is the amount of transmitted data which is caused by higher number of platforms (two vs. one), higher swath (290 km vs. 185 km), higher spectral sampling (13 vs. 11 bands) and higher spatial sampling (4 × 10 m), 6 × 60 m, and 3 × 60 m vs. 1 × 15 m, 8 × 30 m, and 2 × 100 m). These technical improvements represent a new leap in the total amount of freely available earth observation data and hence calls for fast, and easy to parallelize algorithms to allow the efficient processing and exploitation of the incoming data.

In the past, many detection schemes have been developed to detect clouds, cirrus, shadows, snow/ice, and clear sky observations or a subset or superset of these classes. Such schemes are in general distinct for a particular instrument, although basic physical principles for similar spectral bands hold among instruments. Some examples from the relevant literature can be found in a variety of references [1,17–28]. Cloud detection is the main focus for most of these references, while the aim of this study is the separation of all introduced classes.

A comprehensive overview of the existing literature is beyond the scope of this paper, but existing schemes could be characterized by being local or aware of spatial context, self-contained or dependent on external data, or by being probabilistic or decision-based. Local schemes neglect spatial contexts such as texture (e.g., see [25–27]) or objects where for example a cloud shadow could be estimated from the position of a nearby detected cloud, its height, and the given viewing geometry (e.g., see [22]). Self-contained schemes would only depend on the measured data and already available metadata, where other schemes might be based on a time series for this area (e.g., see [1]) or on data from numeric weather prediction models (e.g., see [23]). Probabilistic schemes try to estimate the probability that a given observation belongs to a given class (e.g., see [23,24]). Thus for a given set of classes, the user needs to convert the resulting set of class probabilities into a final decision. It depends on the application if this degree of freedom is welcome or just additional burden. In contrast to this, decision-based schemes select a single class for a given measurement (e.g., see [28]) as the final result. Any probabilistic classification technique can be extended to a decision-based scheme by adding a method which selects a single class from the list of class probabilities as the final result.

Any particular scheme might be a mixture of the discussed approaches or even include strategies not mentioned here. The taken approach will determine not only the classification skill of the algorithm but also the needed effort for implementing and maintaining it as well as the reached processing performance. Although difficult to prove in theory, we assume that local and self-contained approaches are the best choices regarding processing performance and least effort for implementation and maintenance. Such schemes omit the added complexity of processing external dependencies, the computation of spatial metrics and any object recognition and operate only a per-spectrum or per-pixel level. It is of course not guaranteed that such algorithms are inherently fast; e.g., if an online radiative transfer is used for classification. However, this class of algorithms is very well suited for the application of machine learning techniques, which allow rapid development and improvements of algorithms as well as excellent processing speeds. The detection performance of such algorithms can be used to establish a baseline for competing and potentially more sophisticated algorithms to quantitatively assess their potential additional computational costs.

To establish such as baseline, we decided to build up a database of labeled MSI spectra and to apply machine learning techniques to derive ready-to-use classification algorithms. It can serve not only as a valuable tool for algorithm development but also for validation of algorithms. The included spectra should cover much of the natural variability which is seen by MSI while the choice of labels was limited to cloud, cirrus, snow/ice, shadow, water, and clear sky. To our best knowledge, manual classification of images is the most suitable way of setting up such a database for Sentinel-2. To avoid the step of visually inspecting images by a human expert, one could exploit measurements from active instruments such as a LIDAR (e.g., ground-based from EARLINET [29,30] or spaceborne from CALIOP [31]) or cloud radar (e.g., spaceborne from CloudSat [32]). Such measurements should cover large fractions of MSI's swath and potential time delays between data acquisitions should be not greater than several minutes. Currently, suitable space-borne options are not available for Sentinel-2A. Ground-based instruments could be used in principle, but their measurements would cover only a small fraction of the occurring viewing geometries as well as natural variability of surface, atmospheric, and meteorological conditions. The database is discussed in Section 2 and the application of machine learning algorithms is discussed in Section 3.

We understand that the term scheme or technique describes the general method, while a particular, ready-to-use instance of a method with given parameters is an algorithm. The term decision tree refers to the method, while a given tree with branches and parameters is a particular algorithm. We discuss decision trees with features computed from simple band math formulas (see Section 3.2) which are one of the most simple and straightforward to understand techniques. These algorithms are simple to implement for any processing chain, can be represented by simple charts, but offer only limited classification skill. This ease of use and simplicity qualifies decision trees as baseline algorithms to judge the performance of other, possibly more complex and computationally more demanding, algorithms.

To improve the classification skill, we present a detection scheme based on classical Bayesian probability estimation which delivers superior results and is made available to the community as open source software (see Section 3.3). It is a technique that has only recently been used for the detection of clouds [24] and represents a straightforward and fast technique which is very well suited for the processing of large amounts of data. The database, the presented decision trees, as well as the classical Bayesian detection scheme are available at [33].

We included a broad range of available standard methods for our study, but focus the discussion on decision trees and the classical Bayesian approach. In Section 3.4 we discuss these results with regards to other commonly used techniques such as random forests, support vector machines, stochastic gradient descent, and adaptive boosting.

2. Database of Manually Classified Sentinel-2 MSI Data

Sentinel-2 images are selected such that the derived database covers the relevant natural variability of MSI observations. Each included spectrum carries either one of the following labels: cloud, cirrus, snow/ice, shadow, water, and clear sky and is accompanied by the relevant metadata from the Level-1C product. The selection of labels was driven with atmospheric correction of MSI observations in mind, where clear sky, cirrus, and shadow pixels are treated with slightly different approaches. The water class is included for these pixels since remote-sensing-reflectance is a more suitable quantity rather than surface reflectance. Both quantities are products of atmospheric correction algorithms and require different processing steps since for remote-sensing-reflectance the reflection and transmission of the water surface must be corrected. Currently, no external information such as atmospheric fields from numerical weather models like from ERA-interim reanalysis [34] or products from global networks such as aerosol optical depth from Aeronet [35] is included in the database. Since time and location for each included spectra are known, it poses no particular difficulty to extend the database.

The broad global distribution of included images ensures that a wide range of inter-class variability is captured within the database. This variability is highlighted in Figure 1, which depicts histograms of at-sensor reflectances for each MSI channel per class. The figure shows clearly that each of the classes exhibits distinct spectral properties, but also that a classification algorithm has to cope with significant inter- and intra-class variability.

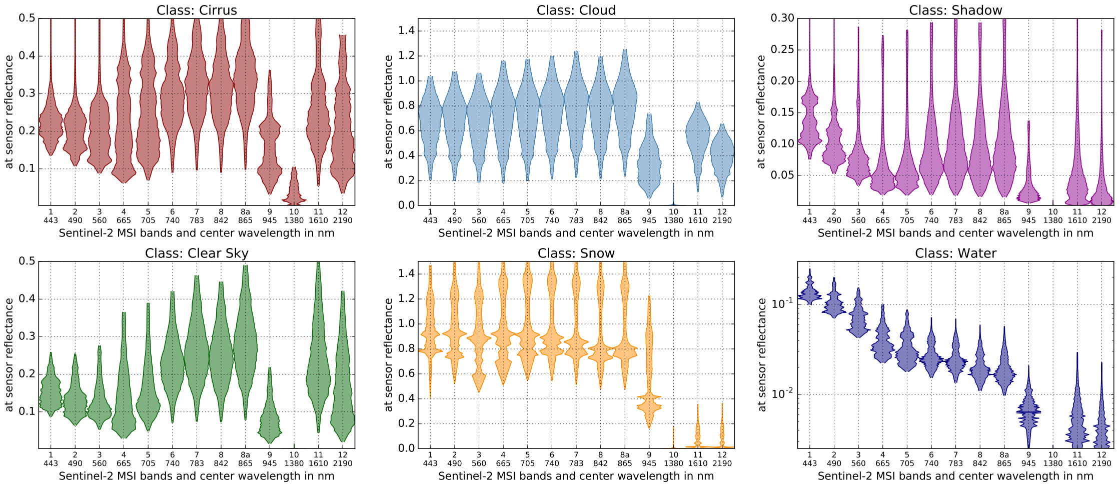


Figure 1. Spectral histograms of the database per class and Sentinel-2 MSI channel. Each panel shows histograms for a single classification class and the data is presented as a Violin plot, where each histogram is normalized to its maximum and is vertically shown on both sides of the baseline line for each channel.

The surface reflectance spectrum and actual atmospheric properties determine spectral shapes of the at-sensor reflectance. This effect can be nicely seen for the so-called cirrus channel B10 at 1380 nm, which shows only small to zero values for most classes other than the cirrus class. The atmosphere is mostly opaque at this wavelength due to the high absorption of water vapor which is mostly concentrated at the first few kilometers above the surface within the planetary boundary layer. Since cirrus clouds typically form well above that height, their scattering properties allow some reflected light to reach the sensor. A different distinct atmospheric band is B9, which is centered at a weaker water vapor absorption band and is used for atmospheric correction. The variability here is caused by surface reflectance and variations in water vapor concentration, which a classification algorithm needs to separate. The reflectance of the snow class decreases substantially for the shortwave infrared channels 11 and 12, while this is not so much the case for the cloud class which shows a mostly flat reflectance spectra in the visible. Also, the increase of surface reflectance at the so-called red edge can be nicely seen in the clear sky class which contains green vegetation.

The database is based on images acquired over the full globe, and their global distribution is illustrated in Figure 2. All spectral bands of the Level-1C products were spatially resampled to 20 m to allow multispectral analysis. The selected scenes are distributed such that a wide range of surface types and observational geometries are included. Multiple false-color RGB views of a scene and it's spatial context is used to label areas with manually drawn polygons. We want to emphasize, that the spatial context of a scene is crucial for its correct manual classification. This is especially the case for shadows cast by clouds or barely visible cirrus. One should also note, that shadows can be cast by objects outside the current image. Not all classes can be found in each image, but we took care to distribute classes evenly.

Figure 3 is a showcase for this manual classification approach for three selected scenes. Their actual positions and used channel combinations are shown in Table 1. It is evident that there is a large degree of freedom on how polygons are placed and which objects are marked. Also, the extent of objects with diffuse boundaries poses a particular burden on the consistency of the manual classification. This merely indicates that a certain degree of subjectivity is inherent in this approach. However, this holds for any other approach when soft objects, such as clouds or shadows, are to be defined by hard boundaries. The effect of the human error is minimized by a two-step approach, where images are initially labeled and revisited some time later to re-evaluate past decisions. Human errors, if present, should lead to unexpected results when comparing labels of the database with classification results from detection algorithms. One example would be prominent misclassification errors between simple-to-distinct classes such as shadow and snow. Next to the spectra itself, the database contains metadata such as the scene-ID, observation time and geographic position, which could be used to establish location-aware algorithms, or to analyze comparison results concerning their location. Currently, metrics for spatial context are not included but could be added with little extra work if needed.

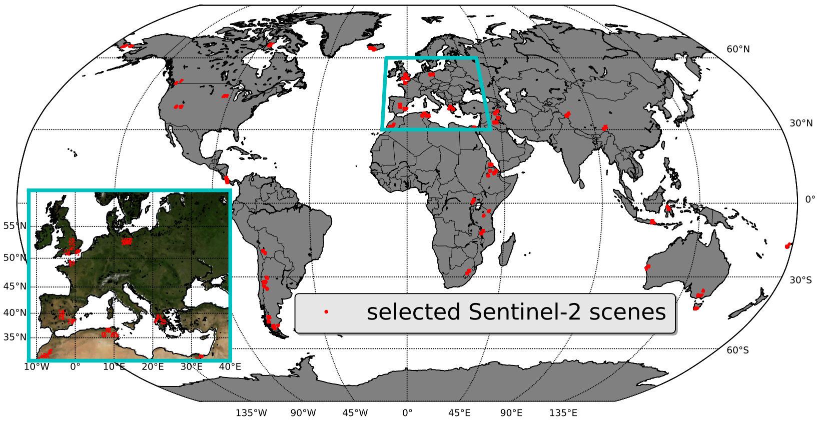


Figure 2. Global distribution of selected Sentinel-2 scenes which are included in the database.

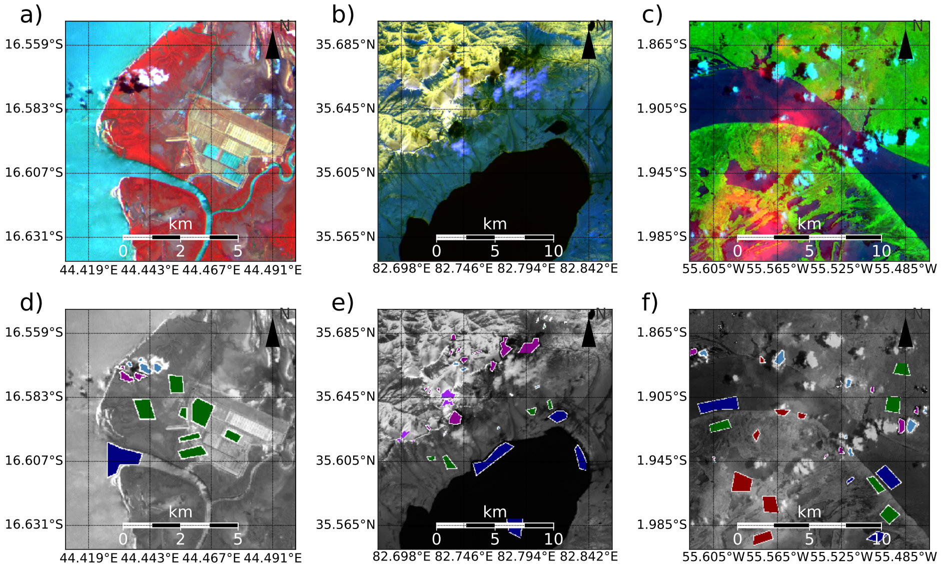


Figure 3. False-color RGB images which have been used to classify Sentinel-2 MSI images manually. The red, green, and blue color channel were composed of appropriate channels (see Table 1) to identify and distinguish classes. The top row of image panels (**a**–**c**) show the complete image, while the bottom row (**d**–**f**) shows manually drawn polygons which identify the various classes. Each polygon border is marked with a white border to simplify their identification within the image. The color of each polygon indicates the class (same colors as in Figure 1 are used: red = cirrus, green = clear sky, dark blue = water, purple = shadow, light blue = cloud) and is consistently used throughout the paper. Additional technical details about the images are given in Table 1.

Table 1. Additional technical information for the images shown in Figure 3. The prefix *S2A_OPER_PRD_MSIL1C_PDMC* was omitted in each given Sentinel-2 product name.

Label	R,G,B	Center Lon	Center Lat	S2 File Name
a,d	8,3,1	44.455°E	16.595°S	20151005T124909_R063_V20151005T072718_20151005T072718
b,e	2,8,10	82.770°E	35.625°N	20151002T122508_R019_V20151002T052652_20151002T052652
c,f	10,7,1	55.545°W	1.925°S	20150928T183132_R110_V20150928T141829_20150928T141829

3. Classification Based on Machine Learning

Labeling Sentinel-2 MSI spectra can be understood as a supervised classification problem in machine learning, for which a rich body of literature and many implemented methods exist. This section introduces shortly some aspects of machine learning which are relevant for this paper and the following subsections provide details about the applied methodology and results. Many methods and available implementations have free parameters which have a substantial impact on the classification result, and optimal parameter settings need to be found for each problem at hand. A common optimization strategy for these parameters is applied here which is described in Section 3.1. This paper is focused on decision trees and classical Bayesian classification, which are discussed in Sections 3.2 and 3.3. Reasons for limiting the study to the two methods is that decision trees are among the most commonly used methods and therefore can establish a baseline and that it is straightforward for both methods to provide algorithms results in a portable way. Portability means that the implementations can run on various hardware and software environments. The classification performance of these two methods is compared with other commonly used methods such as support vector classifiers, random forest, and stochastic gradient descent and results are discussed in Section 3.4.

Supervised classification describes the mapping of input data, which is called feature space, to a fixed and finite set of labels. Here, the feature space is constructed from single MSI spectra using simple band math functions like a single band, band-differences, band-ratios, or generalized indices which are given in Table 2. Only the given functions were considered in this paper, and it is assumed that these functions cover large fractions of the regularly used band relationships. Many suitable techniques have free parameters and it is often necessary to optimize their settings to improve classification results.

Table 2. Band math formulas used for the construction of feature spaces. Names and abbreviations are used throughout this paper.

Name	Short Name	Formula
band	B	$f(a) = a$
difference	S	$f(a,b) = a - b$
ratio	R	$f(a,b) = a/b$
depth	D	$f(a,b,c) = \frac{a+b}{c}$
index	I	$f(a,b) = \frac{a-b}{a+b}$
index$^F_-$	$I+$	$f(a,b,c,d) = \frac{a-b}{c-d}$
index$^F_+$	$I-$	$f(a,b,c,d) = \frac{a+b}{c+d}$

After setup, a classification algorithm can be used as a black box and purely judged by its performance concerning various metrics such as classification skill or needed computational effort. Such an analysis is depicted in Figure 4 for three selected algorithms and two selected features. The features were optimized such that all three algorithms reach similar classification performance. The first column of the figure illustrates results for decision trees which are in principle limited to a rectangular tiling of the feature space. Ready-to-use examples of decision trees are given in Section 3.2. The next column shows results for the same set of transformations, but for a support vector classifier (SVC), which has much more freedom for tiling the feature space. The last column depicts results for the classical Bayesian approach for which a ready-to-use algorithm is discussed in Section 3.3. The separation of the feature space for this algorithm is defined by histograms with respect to a chosen binning scheme and determines the tiling of the feature space. A brief discussion on other possible methods is given in Section 3.4. For visualization purposes, the feature space was limited to two dimensions and only simple transformations based on band math were allowed.

Those examples were chosen to illustrate the difference between selecting a particular machine learning method and selecting appropriate transformations of the input data. These aspects are well known and comprise almost textbook knowledge, but we included this material to highlight the fact that different transformations on the input data lead to different separation of the discussed classes in the feature space. Then, different algorithms can result in various compartmentations of the feature space for the classes, and both choices affect the final classification skill. These problems might have many multiple solutions, of which many can be approximately equivalent for various metrics. Choosing a particular algorithm from such a set of mostly similar algorithms can be random or subjective, without meaningful impact on the final classification result. Discussing individual aspects of a particular algorithm, e.g., a single threshold for a feature on a branch within a decision tree, could be meaningless.

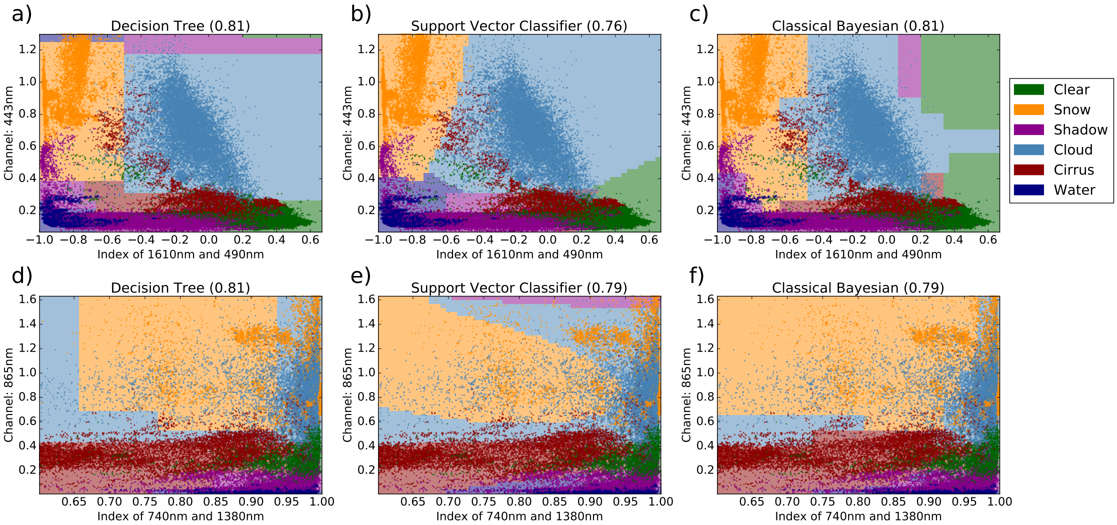


Figure 4. Overview about the behavior of three classification algorithms. Figures in a row ((**a**–**c**) and (**d**–**f**)) are based on the same data transformation (e.g., for the top row, the feature space is build from the 443 nm (B1) channel and an index of the 1610 nm (B11) and 490 nm (B2) channel, where the index function denotes $(\mathrm{i}-\mathrm{j})/(\mathrm{i}+\mathrm{j})$ with i and j are the given channels). Each column (e.g., (**a**,**d**) or (**c**,**f**)) illustrate algorithms based on the same technique which is provided in the title of each plot. The points in each figure show a random sample from the training database and the color indicates the manually defined class. The background color indicates the decision of the algorithm for the full feature space. Indicated in the title of each plot is the ratio of correctly classified samples for the training dataset.

3.1. Optimization and Validation Strategy

All presented classification algorithms are based on a common optimization strategy. We treat the construction of feature spaces and optimization of classification parameters independently and drive the search using random selection techniques. Each step consists then of a randomly selected set of features, a randomly selected set of parameters, and the constructed algorithm. We want to note that many algorithms can reduce the initially given feature space to a smaller, possibly parameterized, number. Feature spaces are constructed using simple band math formulas which are listed in Table 2. The included relations depend on one to four bands, such that a space with n features could depend on up to $4 \times n$ bands.

To separate the training and validation, we randomly split the database into mutually exclusive sets for training and validation. This random split is performed on the total level of the database and includes all datasets as shown in Figure 2. No spectrum which is used for training is therefore used for validation of the algorithm. The classification skill of each algorithm can be evaluated by the ratio of correctly classified cases which naturally varies between zero and one. This score is computed for the training and the validation data set, but only the value of the validation dataset is used for ranking different algorithms. A particular algorithm is excluded if the difference between the two scores becomes too high, which is a simple indication for overfitting.

This understanding of validation is somewhat differently used than in other parts of remote sensing, where validation is usually performed as a comparison of two products (e.g., total column water vapor derived from optical remote sensing measurements and GPS based methods) where one of them is considered as the truth or the product with smaller uncertainty. Here, it is used as confirmation that the machine learning algorithm shows similar results for two separate datasets and that the algorithm doesn't just remember the training dataset.

All presented classification scores are therefore results for the database and the transfer to Sentinel-2 images is based on the assumption that the database is representative of all images. A high value for the classification skill is only a necessary condition for good classification results. If extensive experience with the results of a particular algorithm shows consistency problematic results for specific circumstances, the database should be updated to make it more representative, and the algorithm should be retrained.

3.2. Ready-to-Use Decision Trees

Decision trees can be best understood as a hierarchy of thresholds on single features. All possible decision paths form a tree with each path being a branch. We use the Python library scikit-learn [36] which provides the needed functionality with the CART method [37], which is an optimized version of the C4.5 method [38,39]. It returns optimized decision tree algorithms with prescribed depth for a given training data set which was projected to a selected feature space. The method aims to find a global optimum concerning classification skill and is free of additional parameters.

Decision trees could be constructed manually, but this work can be tedious and doesn't guarantee to deliver better results than automated methods. Both, automated and manual construction aims for a global optimum for given training data and feature space. However, the space of possible decision trees might contain a large number of algorithms with almost equal classification skill near the global optimum. This indicates that the choice of a particular algorithm is certainly not unique and might leave room for discussion. Here, we decided to discuss pragmatically the best trees which we found regarding classification skill for the validation data set. Since the search for optimum feature space is random, we can not guarantee that the global optimum was found, but a long search time was allowed. The search was stopped when after 5000 attempts no better algorithm was found than already known.

Figure 5 depicts a decision tree schematically with depth three and with a ratio of correctly classified spectra of 0.87. The figure also illustrates the success rate of complete separation at the end of each branch. As an example, the final water class contains still a small fraction of shadows, but negligible remainders of other class members. The feature space of this tree is completely composed of bands, and the units are at-sensor reflectance. The ratio of correctly classified spectra increases slightly to 0.89 if band math is allowed within feature space construction. Figure 6 shows the resulting tree in the same style as the previous figure. Both results illustrate that increasing the space of feature spaces can improve the classification performance, but that the effect is not dramatic when decision trees are concerned. This increase in classification skill requires the additional computation of features from bands which should reduce the computational performance of the algorithm.

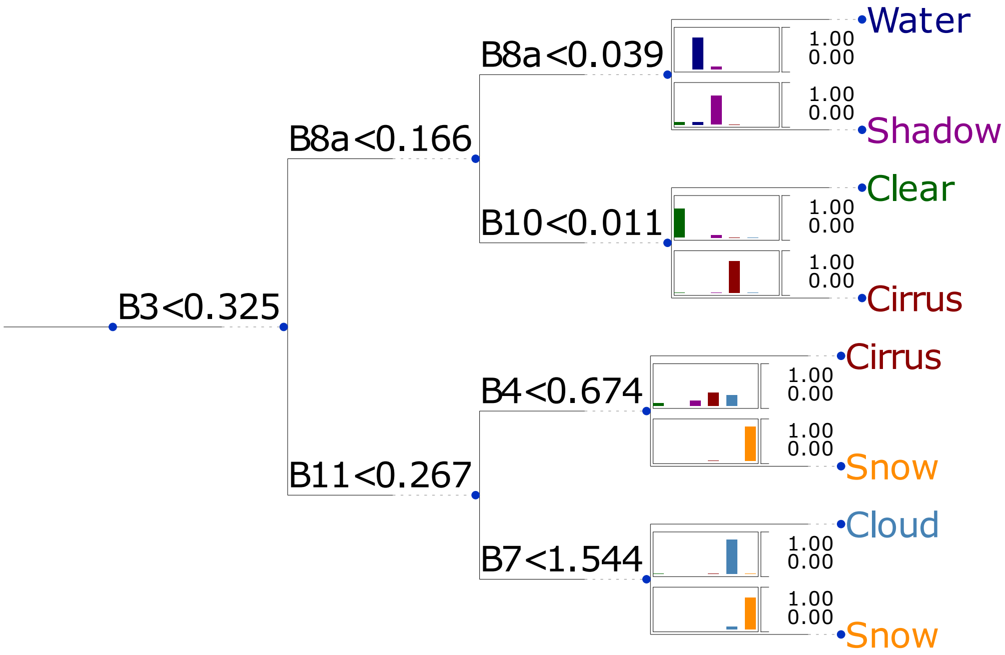


Figure 5. Decision tree of depth three. The tree must be read from left to right, and the actual decision is printed on the horizontal branch. The up direction indicates a *yes* while the down direction indicates a *no* decision. The final class name is shown at the end of each branch. A histogram at the end of each branch indicates the class distribution of samples at the end of the branch and thus indicates the ability of that branch to separate the classes from each other. The ratio of correctly classified spectra of this tree is 0.87.

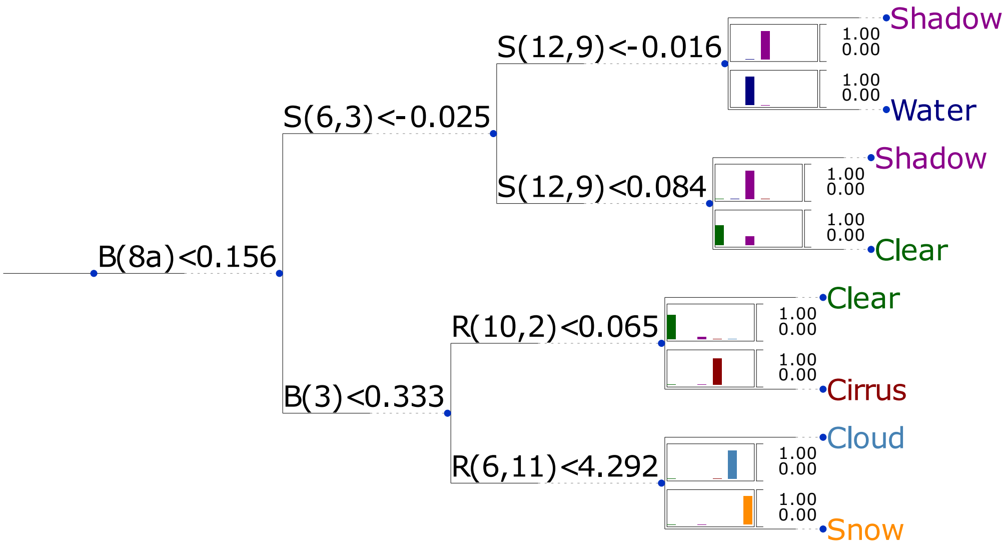


Figure 6. Similar as Figure 5 but this time band math was allowed when the feature space was constructed. Band math functions are defined in Table 2. The ratio of correctly classified spectra of this tree is 0.89.

These result nicely show, that even with a simple technique, a reasonable separation of the classes can be accomplished. The complexity and to some extent the expected classification skill increases with increasing depth of the trees. One can expect, that the increase in classification skill becomes negligible from a particular depth on and that an increase in depth adds only to the complexity and the risk of overfitting. When increasing the allowed depth to four, the ratio of correctly classified spectra increases to 0.91 and a further increase to five only results in a value of 0.92. This indicates that a reasonable regime for classification trees is reached at a depth of four.

Similar as for the case of depth three, results at a depth of four with a feature space composed of bands only is shown in Figure 7, while the result for the constructed feature space is shown in Figure 8. Both algorithms show a similar rate of correctly classified spectra, but the algorithm with the derived feature space shows slightly better results with 0.91 vs. 0.89. It is noteworthy that both decision trees have branches which terminate before the maximum allowed depth is reached. The best solution includes only band math functions for selecting a single band (B), the difference between two bands (S), and the ratio of two bands (R). It is beyond the scope of this work to discuss the physical reasons on how these particular algorithms function. Our focus is to present them in a way which makes them ready-to-use for many applications requiring a pixel mask as input. However, these algorithms are somewhat limited in their classification skill, such that more sophisticated methods might be desirable for certain applications. In the next section, we describe a classification system which reaches a much higher ratio of correctly classified spectra.

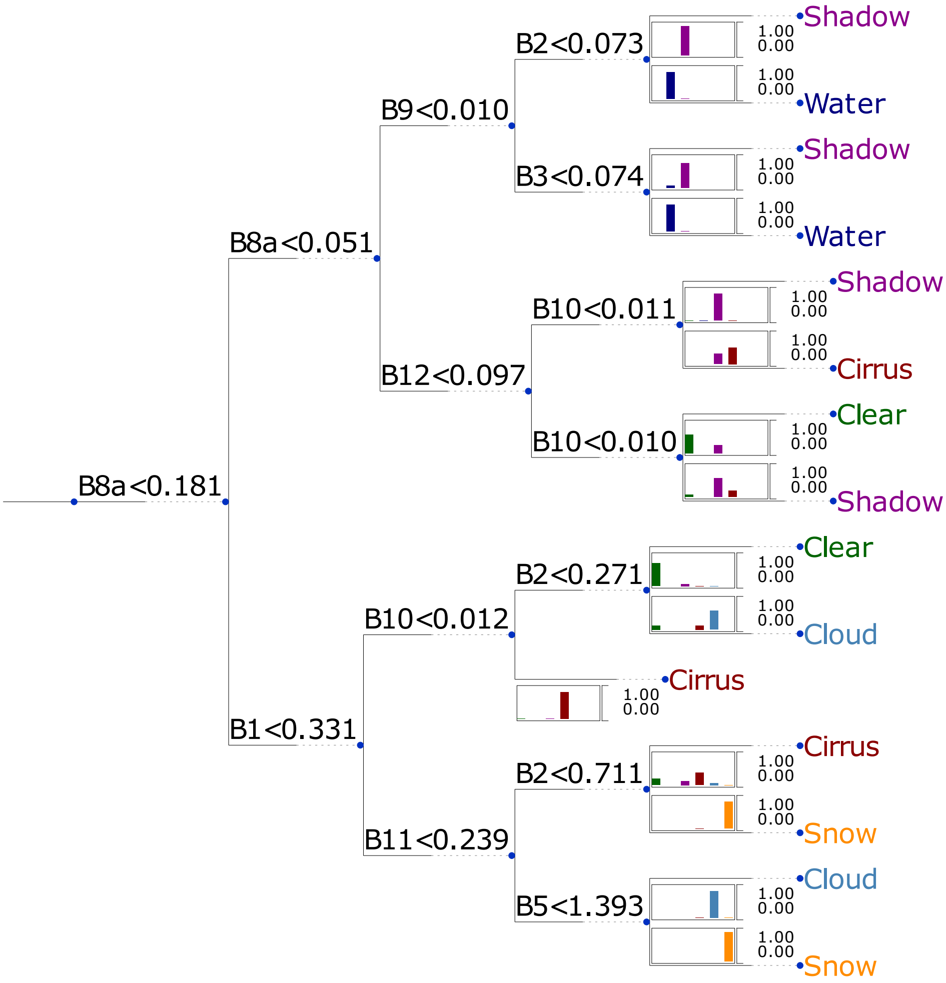


Figure 7. Similar to Figure 5, but for a maximum depth of four. The feature space consists of single bands. For clarity, we omitted the capital B in front of band named when they occur in band math functions. The ratio of correctly classified spectra of this tree is 0.89.

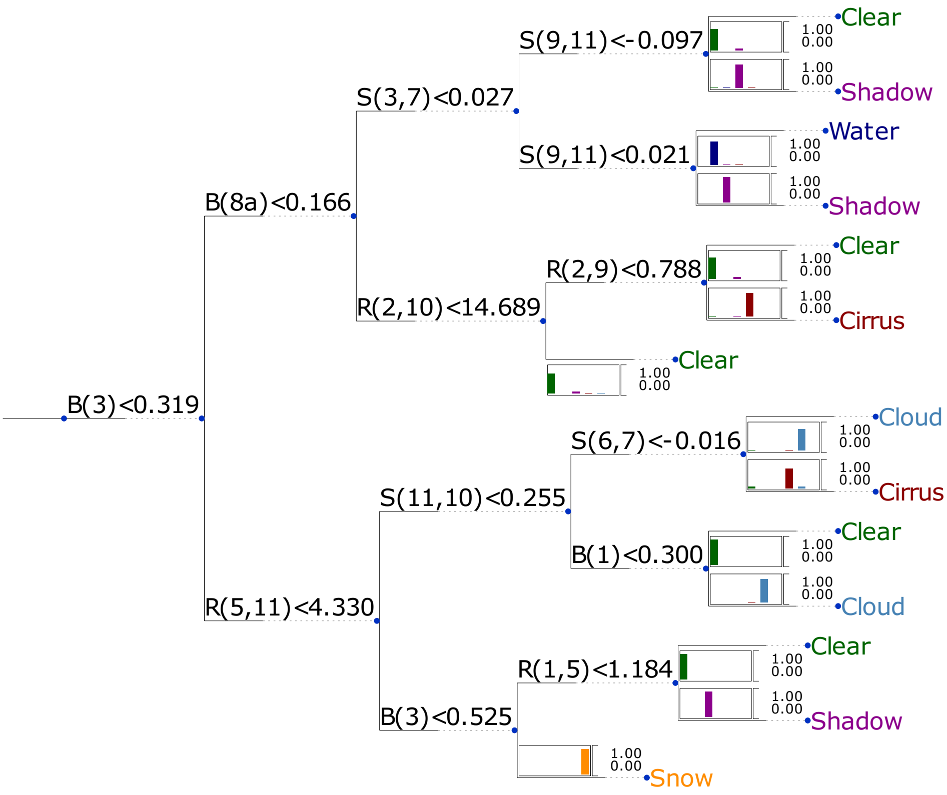


Figure 8. Same as Figure 7, but the feature space is derived from band math functions which are defined in Table 2. The ratio of correctly classified spectra of this tree is 0.91.

3.3. Ready-to-Use Classical Bayesian

Classical Bayesian classification is based on Bayes law for inverting joint probabilities. It can be expressed as:

$$P(C,F) = P(C) \times P(F,C)/P(F) \tag{1}$$

where $P(C,F)$ expresses the joint occurrence probability of class C under the condition of the feature F, with $P(C)$ and $P(F)$ being the global occurrence probabilities of the class C and the feature F. $P(F,C)$ is the joint occurrence probability of the feature F for the class C. The term *classical* distinguishes this approach from the much more commonly used approach of naive Bayesian classification, where one assumes that features are uncorrelated which allows to expresses the joint probability $P(F,C)$ as the product of single occurrence probabilities for each feature. Applications to cloud detection can be found in [23,40,41] while an in-depth discussion of the classical Bayesian approach can be found in [24]. In summary, the joint probability $P(F,C)$ is derived from a histogram of the database with the dimensionality of the number of used features. Free parameters are the number of histogram bins and a smoothing value. The success of this method is based on the selection of the most suitable feature space which follows the previously discussed random approach. The classification is performed by computing the occurrence probability for each class and selecting the one with the highest probability.

Since this method computes occurrence probabilities for each class, it is straightforward to include a confidence measure for each classification. Such a measure can be of great importance in post processing steps, where one might want to process only clear sky pixels for which the classification algorithm was very certain. The construction of such a measure is certainly not unique. We chose to proceed with a simple form, where the sum of all probabilities is normalized to one and the confidence measure is the relative value of the probability of the selected class and the sum of all other probabilities.

Similar as for the case of decision trees, not a single algorithm was found which represents the global optimum, but rather a whole suite of algorithms with similar classification scores. The est found option reaches a ratio of correctly classified spectra of 0.98 for the feature space: $B03 \times S(B9, B1) \times I(B10, B2) \times B12 \times I(B2, B8A)$. Figure 9 shows the confusion matrix for this algorithm, where off-diagonal elements above 0.5% are shown. The confusion matrix was derived from the validation dataset. All classes show excellent values. The largest misclassifications happen between clear sky and shadow and water and shadow classes. This is to be expected for any algorithm since the boundary of the shadow class is diffuse by nature. Also, some smaller confusion between water and shadow should be acceptable since both classes have members who are very dark in all MSI channels.

This figure shows similar information about the classical Bayesian algorithm as the histograms shown for each branch of the discussed decision trees (see Figures 5–8). In contrast to the decision trees, the classification rates are shown for the total classification result and are not broken down for its inner structure. Such an approach is much less straightforward for classical Bayesian algorithms than for decision trees since the exploited features are used at once and are not ordered within an internal hierarchy.

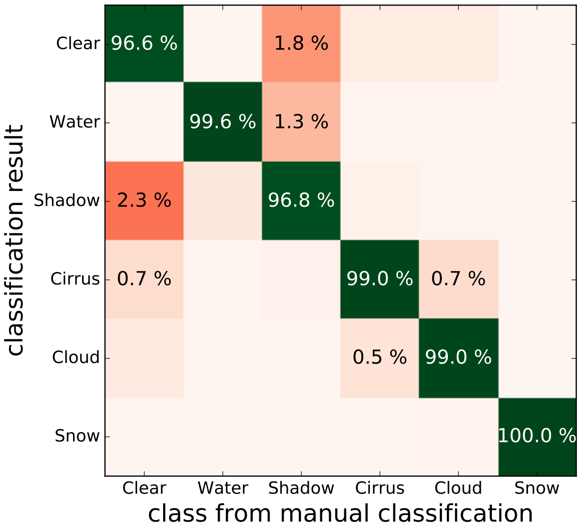
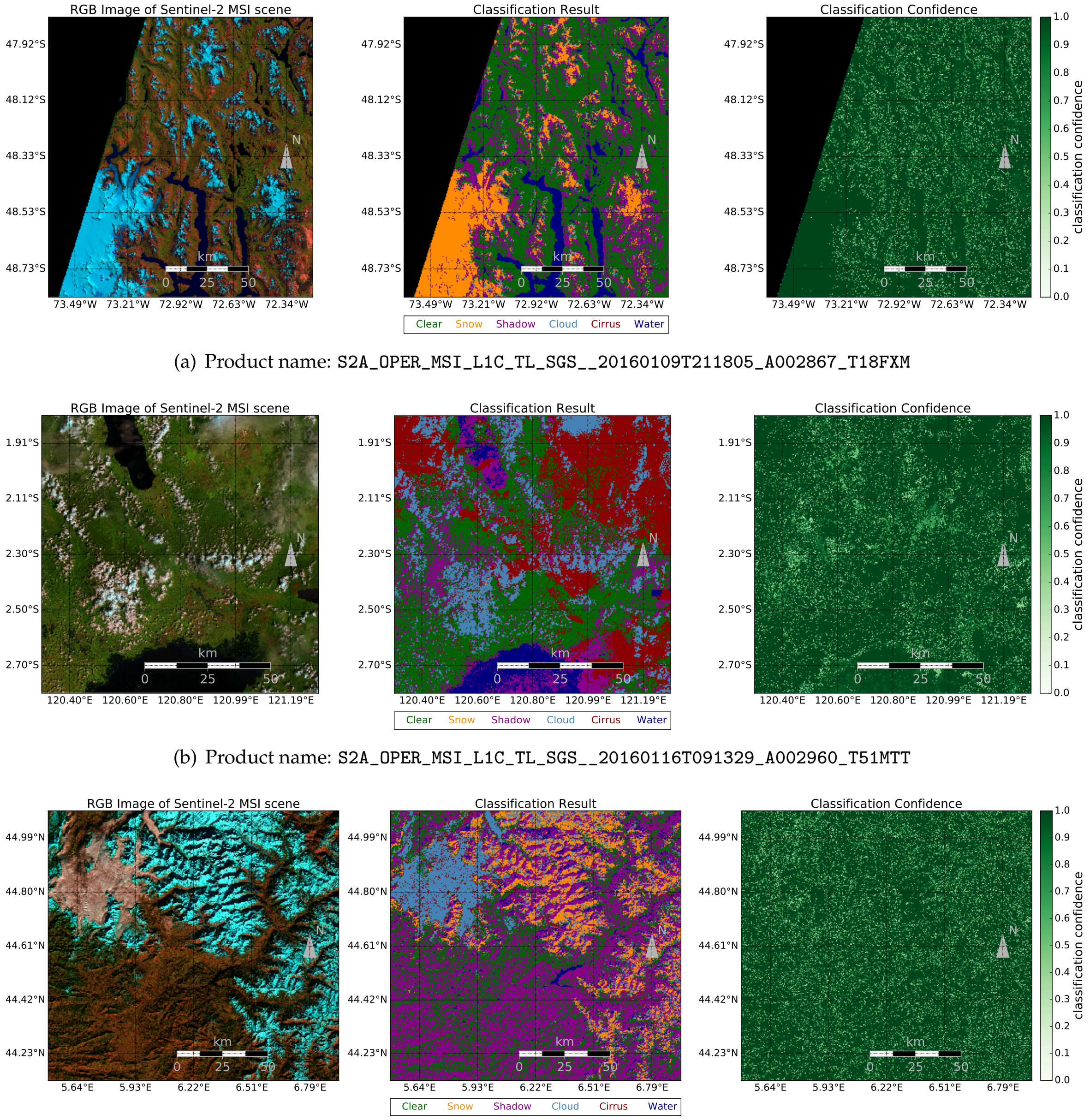


Figure 9. Overview about the rate of classification for each class concerning all other classes. Apart from round-off errors, data adds up to 100% column-wise for each class, but only values above 0.5% are shown to increase the clarity of the figure. The data sample is based on manually classified data which were not used for setting up the algorithm.

Figure 10 shows RGB images of Sentinel-2 scenes together with the derived mask and the classification confidence. The figure captures diverse areas and includes mountainous regions, green vegetation, snow and ice, as well as clouds and shadows. Especially the mountainous regions show that a separation of ice and shadows is achieved. Many pixels are marked as affected by shadows, which can be useful when atmospheric correction is performed. Only a few spatial patterns from the images are present in the confidence masks which indicates that this measure is mostly independent of the scene. Some larger water bodies and snow-covered areas can be found in the confidence maps as areas of reduced noise, but this only illustrates the homogeneity of the scene itself. The used color scale does not exaggerate small variations of the confidence since these might be hard to interpret. However, depending on the actual application, it can be straightforward to use the confidence value and together with appropriately selected thresholds to filter data for further processing. The dark water bodies in panel b of the figure include some areas which are wrongly classified as shadow. This can be expected from the analysis of the confusion matrix (see Figure 9). The confidence value for these areas is only slightly reduced.



(a) Product name: `S2A_OPER_MSI_L1C_TL_SGS__20160109T211805_A002867_T18FXM`

(b) Product name: `S2A_OPER_MSI_L1C_TL_SGS__20160116T091329_A002960_T51MTT`

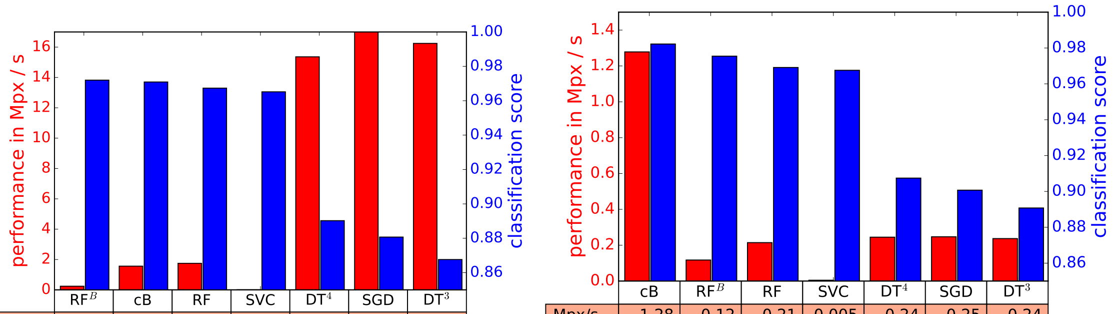
(c) Product name: `S2A_OPER_MSI_L1C_TL_SGS__20160205T174515_A003251_T31TGK`

Figure 10. (**a**–**c**) Overview of classification results based on the classical Bayesian algorithm. The name of the used Level-1C product is given in the caption below each panel. Within a figure, the left panel shows an RGB (B11,B8,B3) view of the scene, the middle panel shows the classification mask, and the right panel shows the classification confidence.

3.4. Comparison with Commonly Used Techniques

This work focuses on decision trees and classical Bayesian algorithms, but the results from other techniques help to put them into perspective. Commonly used techniques from the domain of machine learning are random forests (RF) [42–44], support vector classifiers (SVC) [45], and stochastic gradient descent (SGD) [46,47]. Methods which also compute class probabilities can be combined using adaptive boosting [48,49]. We used the implementations of these techniques in scikit-learn to derive additional classification algorithms using the same random search approach as outlined above.

Before presenting the results, we want to emphasize that they represent merely lower boundaries for classification skill and processing speed. The used techniques have free parameters which were sampled by a random search, but experts for particular techniques might be able to use better implementations or find better sets of parameters and hence produce better or faster results than the ones presented here. However, the results might be representative for a typical user who uses an implementation, but is not aware of all possible details. An overview about the numeric experiments is given in Figure 11. The analysis was separated by the types of allowed feature spaces. The left panel shows results for original bands only, while for the right panel feature spaces based on band math based on Table 2 were allowed. All presented decision trees from Figure 5 to Figure 8 are included as well as the classical Bayesian algorithm which was discussed in Section 3.3. The classical Bayesian for the single band feature space is based on $B1 \times B4 \times B8 \times B10 \times B11$. The parameters of the other techniques are not listed here since they depend on a specific implementation and might not be portable among different implementations.



	RF^B	cB	RF	SVC	DT^4	SGD	DT^3
Mpx/s	0.24	1.56	1.75	0.002	15.36	17.09	16.25
Score:	0.972	0.971	0.967	0.965	0.890	0.881	0.868
Shadow	0.94	0.94	0.94	0.93	0.76	0.76	0.72
Water	0.99	1.00	0.99	0.99	0.93	0.95	0.90
Snow	1.00	1.00	1.00	0.99	0.97	1.00	0.99
Cirrus	0.97	0.98	0.97	0.95	0.86	0.88	0.86
Cloud	0.99	0.99	0.98	0.99	0.95	0.94	0.92
Clear	0.96	0.95	0.95	0.95	0.91	0.84	0.88

(a) Feature space based on single bands only.

	cB	RF^B	RF	SVC	DT^4	SGD	DT^3
Mpx/s	1.28	0.12	0.21	0.005	0.24	0.25	0.24
Score:	0.982	0.975	0.969	0.968	0.907	0.901	0.891
Shadow	0.97	0.95	0.94	0.93	0.77	0.87	0.75
Water	1.00	0.99	0.99	0.99	0.98	0.98	0.97
Snow	1.00	1.00	1.00	1.00	0.99	0.99	1.00
Cirrus	0.99	0.98	0.97	0.97	0.87	0.88	0.83
Cloud	0.99	0.99	0.99	0.99	0.93	0.96	0.95
Clear	0.97	0.97	0.96	0.95	0.93	0.80	0.90

(b) Feature based on band math functions as defined in Table 2

Figure 11. Overview about global and per-class classification skill (blue bars, right scale) and algorithm performance in Mega Pixel per second (Mpx/s, red bars, left scale) for different machine learning techniques. Panel (**a**) shows results for feature spaces based on single bands while panel (**b**) shows results for feature spaces based on band math functions. The classification scale of both panels is the same for better comparison. Tables below the two panels report numbers for classification performance as well as the classification score for the full validation dataset and separated by individual classes. The labels are: *RF* = Random Forest (*RF*^{B} includes adaptive boosting), *cB* = classical Bayesian, *SVC* = support vector classifiers, *SGD* = stochastic gradient descent, *DT*^{n} = decision tree of depth *n*. The algorithms are sorted by their classification skill.

The processing performance in Mega Pixel per second (Mpx/s) is also reported. All computations were performed on an Intel i5-3570 CPU @ 3.40 GHz. These figures depend on the implementation of a particular technique and should only be used to judge a particular implementation and not the technique as such. The classical Bayesian is implemented in the Python language and uses parts of NumPy and SciPy [50]. It seems obvious that algorithms which use a feature space based purely on bands should have an advantage since the calculation of features can be skipped. This can be nicely seen for the classical Bayesian algorithms, but breaks for the decision trees as well as the stochastic gradient descent, which only shows that even for a single implementation of a technique the real run-times can vary drastically. Possible reasons for this effect might be a change of the used numeric type between the two approaches.

When the classification skill is concerned, random forests, support vector classifiers, and the classical Bayesian show quite similar results and other factors should be discussed when a particular algorithm needs to be selected. Although not impossible, portability of scikit-learn algorithms is currently problematic and only safely possible when the algorithm is retrained on each new computer system. Classical Bayesian algorithms are safely portable among setups which provide a recent python distribution. Also, it shows good processing performance even when implemented in a language which is sometimes referred to as slow.

For the scikit-learn techniques, adaptive boosting was only applied to the random forests since other methods did not provide class probabilities. Only minor improvements are found, which can be expected since random forest already is an ensemble method. The classical Bayesian provides probabilities, and adaptive boosting can be applied. It is not discussed here since it was planned to share the algorithm and a combination with scikit-learn introduces problems with portability.

The classification scores on a per-class level show expected results with the smallest results for shadow and clear sky pixels. This is caused by the diffuse nature of both classes and should not come as a surprise. All in all, the classes show all good scores and the discussion of confusion matrices is not needed.

The classical Bayesian gains little classification skill for feature spaces based on band math. In general, this method is limited to smaller numbers of features and was limited to five for this study. The number of selected features defines the dimensionality of the underlying histograms, and the number of bins sets requirements for needed computer memory as well as the amount of required training data. This indicates that this limit is more practical than theoretical. The stability of the classification skill for both approaches shows that all relevant information for this task can be included within five features.

4. Conclusions

A database of manually labeled spectral from Sentinel-2 MSI was set-up and presented. It contains spectra as well as metadata such as observational geometry and geographic position. The data is labeled with the classes: shadow, snow, cirrus, cloud, water, and clear sky and can be used to create and to validate classification algorithms. The considered classes are crucial for atmospheric correction as well as other methods which rely on pixel masks for the filtering of input data.

The series of Sentinel-2 platforms will deliver unprecedented amounts of Earth observation data which calls for fast and efficient algorithms for data processing. Such algorithms can establish a baseline to quantitatively evaluate the added value of algorithms with a higher demand on computational resources. Machine learning techniques offer straightforward routes for the development of fast algorithms and were applied to derive ready-to-use classification algorithms. Decision trees were discussed since they are simple-to-understand and easy-to-implement and are therefore suitable candidates for baseline algorithms. It was found that trees of depth four show a classification performance as good as trees with higher depth. The ratio of correctly classified spectra reached 0.91, while trees of depth three can reach values of 0.87. Several ready-to-use decision trees were presented in schematic form. Feature spaces based on the bands alone as well as based on band math were tested. For algorithms with the same number of features, those based on the full range of band math formulas gave only slightly better results than those using single bands only. An algorithm based on the classical Bayesian approach was discussed to increase the classification performance to a value of 0.98. A limiting factor in a further increase of the detection skill is the inherent diffuse separation of clear sky and shadow pixels. Smaller effects are caused by a misclassification of dark water and shadow pixels. The discussed database, the presented decision trees, as well as the classical Bayesian approach are available at [33].

A comparison with other widely used machine learning techniques shows, that similar results can be achieved with random forests and support vector classifiers. Since portability and processing performance can be an issue, the classical Bayesian algorithm is a good candidate for general use and distribution of algorithms.

The presented classification is an essential pre-processing step, which in most workflows will be followed by additional processing or further classification. The derived classification can then be used as input data filter for these steps.

Selecting an actual algorithm should be based on user requirements which diminish the value of general suggestions. If users need full control over the algorithm and are willing to invest labor, the database and one of the suggested machine learning methods can be applied. This approach is of particular value if the detection of a subset of the presented class is required with higher accuracy than others. If only a minimum of extra work can be spent on this task, potential users should choose either the presented implementation based on the classical Bayesian or one of the presented decision trees. Selection of the decision trees depends mainly on the required processing speeds and classification performance. For highest processing speeds, the decision tree of depth three based on single bands should be used (see Figure 5). If this requirement can be relaxed, the decision tree of depth four based on band math (see Figure 8) delivers better classification performance with possibly decreased processing speeds. Much better classification performance can be expected from the classical Bayesian implementation which is ready-to-use.

Acknowledgments: We want to acknowledge ESA for providing access to early data products Sentinel-2.

Author Contributions: André Hollstein wrote the manuscript, implemented the classical Bayesian method, and performed the numeric experiments and analyzed their results. Karl Segl and Maximilian Brell helped with reading and interpreting early Sentinel-2 data. Luis Guanter, Karl Segl, and Maximilian Brell contributed to the analysis of the results and the outline of the manuscript. Marta Enesco was responsible for the manual classification of Sentinel-2 data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hagolle, O.; Huc, M.; Pascual, D.V.; Dedieu, G. A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENμS, LANDSAT and SENTINEL-2 images. *Remote Sens. Environ.* **2010**, *114*, 1747–1755.
2. Muller-Wilm, U.; Louis, J.; Richter, R.; Gascon, F.; Niezette, M. Sentinel-2 level 2A prototype processor: Architecture, algorithms and first results. In Proceedings of the ESA Living Planet Symposium, Edinburgh, UK, 9–13 September 2013.
3. Yan, L.; Roy, D.P.; Zhang, H.; Li, J.; Huang, H. An automated approach for sub-pixel registration of Landsat-8 Operational Land Imager (OLI) and Sentinel-2 Multi Spectral Instrument (MSI) imagery. *Remote Sens.* **2016**, *8*, 520.
4. Reddy, B.S.; Chatterji, B.N. An FFT-based technique for translation, rotation, and scale-invariant image registration. *IEEE Trans. Image Process.* **1996**, *5*, 1266–1271.
5. Paul, F.; Winsvold, S.H.; Kääb, A.; Nagler, T.; Schwaizer, G. Glacier remote sensing using Sentinel-2. Part II: Mapping glacier extents and surface facies, and comparison to Landsat 8. *Remote Sens.* **2016**, *8*, 575.
6. Du, Y.; Zhang, Y.; Ling, F.; Wang, Q.; Li, W.; Li, X. Water bodies' mapping from Sentinel-2 imagery with modified normalized difference water index at 10-m spatial resolution produced by sharpening the SWIR band. *Remote Sens.* **2016**, *8*, 354.
7. Lefebvre, A.; Sannier, C.; Corpetti, T. Monitoring urban areas with Sentinel-2A data: Application to the update of the Copernicus high resolution layer imperviousness degree. *Remote Sens.* **2016**, *8*, 606.
8. Pesaresi, M.; Corbane, C.; Julea, A.; Florczyk, A.J.; Syrris, V.; Soille, P. Assessment of the Added-Value of Sentinel-2 for Detecting Built-up Areas. *Remote Sens.* **2016**, *8*, 299.
9. Immitzer, M.; Vuolo, F.; Atzberger, C. First experience with Sentinel-2 data for crop and tree species classifications in Central Europe. *Remote Sens.* **2016**, *8*, 166.
10. Clevers, J.G.; Gitelson, A.A. Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on Sentinel-2 and -3. *Int. J. Appl. Earth Obs. Geoinform.* **2013**, *23*, 344–351.
11. Martimor, P.; Arino, O.; Berger, M.; Biasutti, R.; Carnicero, B.; Del Bello, U.; Fernandez, V.; Gascon, F.; Silvestrin, P.; Spoto, F.; et al. Sentinel-2 optical high resolution mission for GMES operational services. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Barcelona, Spain, 23–28 July 2007; pp. 2677–2680.
12. Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; et al. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sens. Environ.* **2012**, *120*, 25–36.

13. Malenovský, Z.; Rott, H.; Cihlar, J.; Schaepman, M.E.; García-Santos, G.; Fernandes, R.; Berger, M. Sentinels for science: Potential of Sentinel-1,-2, and-3 missions for scientific observations of ocean, cryosphere, and land. *Remote Sens. Environ.* **2012**, *120*, 91–101.
14. Irons, J.R.; Dwyer, J.L.; Barsi, J.A. The next Landsat satellite: The Landsat data continuity mission. *Remote Sens. Environ.* **2012**, *122*, 11–21.
15. Hansen, M.C.; Loveland, T.R. A review of large area monitoring of land cover change using Landsat data. *Remote Sens. Environ.* **2012**, *122*, 66–74.
16. Roy, D.P.; Wulder, M.; Loveland, T.; Woodcock, C.; Allen, R.; Anderson, M.; Helder, D.; Irons, J.; Johnson, D.; Kennedy, R.; et al. Landsat-8: Science and product vision for terrestrial global change research. *Remote Sens. Environ.* **2014**, *145*, 154–172.
17. Rossow, W.B.; Garder, L.C. Cloud detection using satellite measurements of infrared and visible radiances for ISCCP. *J. Clim.* **1993**, *6*, 2341–2369.
18. English, S.; Eyre, J.; Smith, J. A cloud-detection scheme for use with satellite sounding radiances in the context of data assimilation for numerical weather prediction. *Q. J. R. Meteorol. Soc.* **1999**, *125*, 2359–2378.
19. Choi, H.; Bindschadler, R. Cloud detection in Landsat imagery of ice sheets using shadow matching technique and automatic normalized difference snow index threshold value decision. *Remote Sens. Environ.* **2004**, *91*, 237–242.
20. Gómez-Chova, L.; Camps-Valls, G.; Amorós-López, J.; Guanter, L.; Alonso, L.; Calpe, J.; Moreno, J. New cloud detection algorithm for multispectral and hyperspectral images: Application to ENVISAT/MERIS and PROBA/CHRIS sensors. In Proceedings of the IEEE International Symposium on Geoscience and Remote Sensing, Denver, CO, USA, 31 July–4 August 2006; pp. 2757–2760.
21. Zhu, Z.; Woodcock, C.E. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sens. Environ.* **2012**, *118*, 83–94.
22. Zhu, Z.; Wang, S.; Woodcock, C.E. Improvement and expansion of the Fmask algorithm: Cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images. *Remote Sens. Environ.* **2015**, *159*, 269–277.
23. Heidinger, A.K.; Evan, A.T.; Foster, M.J.; Walther, A. A naive Bayesian cloud-detection scheme derived from CALIPSO and applied within PATMOS-x. *J. Appl. Meteorol. Climatol.* **2012**, *51*, 1129–1144.
24. Hollstein, A.; Fischer, J.; Carbajal Henken, C.; Preusker, R. Bayesian cloud detection for MERIS, AATSR, and their combination. *Atmos. Meas. Tech.* **2015**, *8*, 1757–1771.
25. Merchant, C.; Harris, A.; Maturi, E.; MacCallum, S. Probabilistic physically based cloud screening of satellite infrared imagery for operational sea surface temperature retrieval. *Q. J. R. Meteorol. Soc.* **2005**, *131*, 2735–2755.
26. Visa, A.; Valkealahti, K.; Simula, O. Cloud detection based on texture segmentation by neural network methods. In Proceedings of the IEEE International Joint Conference on Neural Networks, Singapore, 18–21 November 1991; pp. 1001–1006.
27. Song, X.; Liu, Z.; Zhao, Y. Cloud detection and analysis of MODIS image. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Anchorage, AK, USA, 20–24 September 2004; Volume 4, pp. 2764–2767.
28. Derrien, M.; Farki, B.; Harang, L.; LeGleau, H.; Noyalet, A.; Pochic, D.; Sairouni, A. Automatic cloud detection applied to NOAA-11/AVHRR imagery. *Remote Sens. Environ.* **1993**, *46*, 246–267.
29. Matthais, V.; Freudenthaler, V.; Amodeo, A.; Balin, I.; Balis, D.; Bösenberg, J.; Chaikovsky, A.; Chourdakis, G.; Comeron, A.; Delaval, A.; et al. Aerosol lidar intercomparison in the framework of the EARLINET project. 1. Instruments. *Appl. Opt.* **2004**, *43*, 961–976.
30. Amodeo, A.; Pappalardo, G.; Bösenberg, J.; Ansmann, A.; Apituley, A.; Alados-Arboledas, L.; Balis, D.; Böckmann, C.; Chaikovsky, A.; Comeron, A.; et al. A European research infrastructure for the aesorol study on a continental scale: EARLINET-ASOS. *Proc. SPIE* **2007**, doi:10.1117/12.738401.
31. Winker, D.M.; Vaughan, M.A.; Omar, A.; Hu, Y.; Powell, K.A.; Liu, Z.; Hunt, W.H.; Young, S.A. Overview of the CALIPSO Mission and CALIOP Data Processing Algorithms. *J. Atmos. Ocean. Technol.* **2009**, *26*, 2310–2323.
32. Stephens, G.L.; Vane, D.G.; Boain, R.J.; Mace, G.G.; Sassen, K.; Wang, Z.; Illingworth, A.J.; O'Connor, E.J.; Rossow, W.B.; Durden, S.L.; et al. The CloudSat mission and the A-Train: A new dimension of space-based observations of clouds and precipitation. *Bull. Am. Meteorol. Soc.* **2002**, *83*, 1771–1790.
33. Hollstein, A. Classical Bayesian for Sentinel-2. Available online: https://github.com/hollstein/cB4S2 (accessed on 16 August 2016).

34. Dee, D.; Uppala, S.; Simmons, A.; Berrisford, P.; Poli, P.; Kobayashi, S.; Andrae, U.; Balmaseda, M.; Balsamo, G.; Bauer, P.; et al. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **2011**, *137*, 553–597.
35. Holben, B.; Eck, T.; Slutsker, I.; Tanré, D.; Buis, J.; Setzer, A.; Vermote, E.; Reagan, J.; Kaufman, Y.; Nakajima, T.; Lavenu, F.; Jankowiak, I.; Smirnov, A. AERONET—A federated instrument network and data archive for aerosol characterization. *Remote Sens. Environ.* **1998**, *66*, 1–16.
36. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
37. Lewis, R.J. An introduction to classification and regression tree (CART) analysis. In Proceedings of the Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, San Francisco, CA, USA, 23 May 2000; pp. 1–14.
38. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann: Burlington, MA, USA, 1993; Volume 1.
39. Quinlan, J.R. Improved use of continuous attributes in C4.5. *J. Artif. Intell. Res.* **1996**, *4*, 77–90.
40. Mackie, S.; Embury, O.; Old, C.; Merchant, C.; Francis, P. Generalized Bayesian cloud detection for satellite imagery. Part 1: Technique and validation for night-time imagery over land and sea. *Int. J. Remote Sens.* **2010**, *31*, 2573–2594.
41. Uddstrom, M.J.; Gray, W.R.; Murphy, R.; Oien, N.A.; Murray, T. A Bayesian cloud mask for sea surface temperature retrieval. *J. Atmos. Ocean. Technol.* **1999**, *16*, 117–132.
42. Ho, T.K. Random decision forests. In Proceedings of the IEEE Third International Conference on Document Analysis and Recognition, Washington, DC, USA, 14 –15 August 1995; Volume 1, pp. 278–282.
43. Ho, T.K. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 832–844.
44. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
45. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297.
46. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2005**, *67*, 301–320.
47. Xu, W. Towards optimal one pass large scale learning with averaged stochastic gradient descent. *arXiv preprint* **2011**, arXiv:1107.2490.
48. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139.
49. Zhu, J.; Zou, H.; Rosset, S.; Hastie, T. Multi-class adaboost. *Stat. Interface* **2009**, *2*, 349–360.
50. Van der Walt, S.; Colbert, S.; Varoquaux, G. The NumPy Array: A structure for efficient numerical computation. *Comput. Sci. Eng.* **2011**, *13*, 22–30.